The Reference Model is the Most Validated Diabetes Cardiovascular Model Known

by: Jacob Barhak

2019 IMAG Multiscale Modeling Consortium (MSM) Meeting 6-7 March 2019

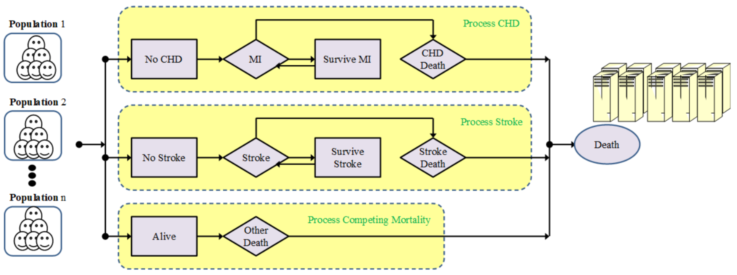
The Reference Model accumulates knowledge from many models and observed outcomes imported from ClinicalTrials.Gov - It now validates against more populations than any other known model!

Abstract:

The Reference Model is an ensemble model that accumulates knowledge from multiple other models and validates this knowledge against multiple populations. After connecting to ClinicalTrials.Gov it has been growing rapidly and has reached the point where it validates against more populations than any other known diabetes model. It currently contains 30 risk models that cooperate and compete and assemble the best model that fits 123 cohorts from 31 populations. This year there was an increase in the number of cohort downloaded from ClinicalTrials.Gov, yet more importantly the cumulative computational gap of knowledge can now be explored interactively via the web. This gap of knowledge shows the difference between the model predication and the results for each clinical trial cohort. The Reference Model accumulates models and data, so being able to show this gap for accumulated knowledge represents our limits to model diabetes. If ClinicalTrials.Gov would be better standardized, it would be even easier to import data. More data can help narrow this gap that can now be calculated and visualized. Such improvements may lead in the long run for better models that may be used for decision making now reserved for humans. However, to allow such advanced modeling, ClinicalTrials.Gov data requires standardization.

The Reference Model

- Ensemble model
- Accumulates knowledge from:
  - Existing models
  - Observed outcomes
- Focuses on summary data
  - Avoids individual data restrictions
  - Larger merged population base
- Flexible Import from ClinicalTrials.Gov
- Applicable for other disease processes
- Traceable and reproducible
- Can map our understanding gap
- Currently focuses on diabetic populations
- Output is now available online through this poster
- Protected by U.S. Patent 9,858,390



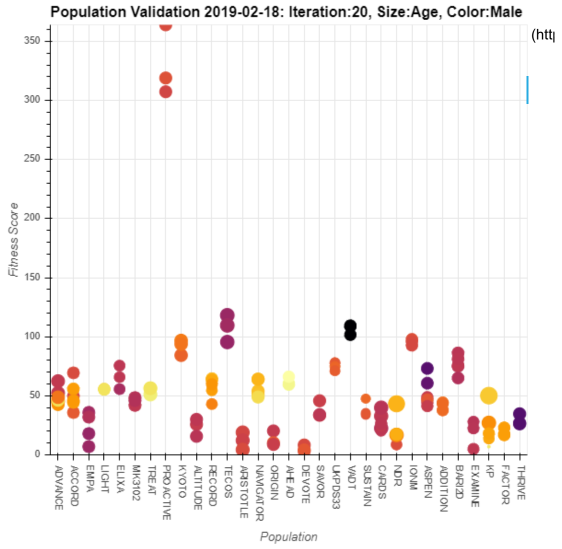
What is new this year?

- The model imported new populations this year from ClinicalTrials.Gov.
  - The number of populations/cohorts increased this year from 21 with 91 cohorts to 31 with 123 cohorts.
  - This number of cohorts is far beyond the previous largest validation of a Diabetes model published - the Archimedes model which reported 18 trials with 74 exercises.
- Visualization was improved and now interactive visualization of results
  - Users can explore more parameters by hovering over populations
  - Users can explore the populations by changing size of marker and color of marker using predetermined selections
  - Users can explore the optimization of mixture of models using a slider

Results:

Below are simulation results. The top plot shows the gap of our cumulative computational understanding by showing modeling error for cohorts of clinical studies. Each column of circles represents a different study. The vertical axis is the fitness score calculated for the model mixture at a specific iteration. Each circle represents a cohort. The circles at the bottom represent populations that the models used explain better than others. Circles near the top of the plot represent outliers that the current model cannot explain well. The model mixture is explained at the bottom plot that shows the model mixture during the optimization process that calculates the best model mixture. The circle at the left represent the overall fitness being optimized. Each bar represents the contribution of each model. Models of the same color represent different risk equations and assumptions that cooperate during optimization to explain the same phenomenon. The iteration slider allows exploring the model mixture. When moving the slider, one will see some models loose influence and other gain influence. The best model mixture is seen when the iteration slider is at the highest number.

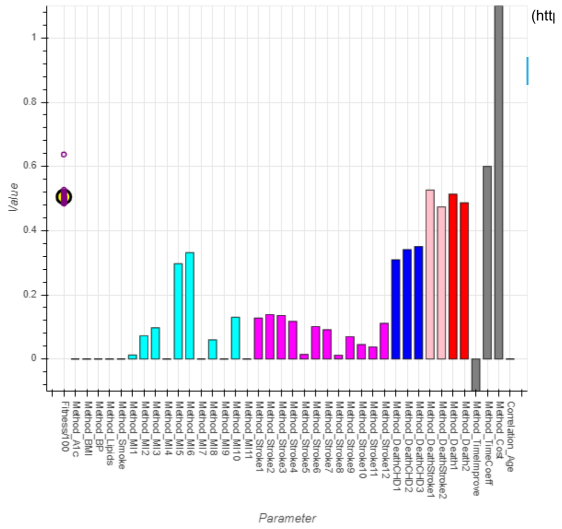
Population Validation 2019-02-18: Iteration:20, Size:Age, Color:Male



Display: Size: Age, Color: Male ▾

Iteration: 20

Model Mixture 2019-02-18: Iteration: 20



Iteration: 20

Discussion and Future Efforts:

Although there are visible outliers, the term validation in the context of this model is correct. The Reference Model is a validation model aimed at validating multiple models against as much clinical data as can be accumulated. The fitness score represents a calculated difference between multiple modeled and observed outcomes for 1000 individuals. The average fitness reached in this simulation is 50/1000 which is larger than the 32/1000 published last year - indicating that the 10 populations added this year worsened our computational comprehension gap. It means we need better explanations. The outlier population seen in the plot is most probably a modeling error related to misinterpretation of outcomes in the PROactive study follow up clinical trial. The trial has only one outcome reported aggregating many other outcomes which requires interpretation. Such misinterpretation will be common since clinical trial reports are still not standardized and therefore there is much room for expert interpretation and the data is not yet computer comprehensible. Some efforts are planned to improve such modeling capabilities by: *Incorporating human expert interpretation* Standardizing clinical trial data - see the accompanying poster for initial steps * importing more data from ClinicalTrials.Gov - with the fast growth of this database, accumulation of knowledge is easier than ever before.

Technology:

The Python programming language is the main technological enabler behind the model. The new visualization through a web browser is possible using the holoviews library that allows plotting and user interaction with the data. The Reference Model itself runs simulations using the MIcro Simulation Tool (MIST) that runs simulations in parallel on multiple machines on multiple CPUs . It is possible to run those simulations on the Amazon Elastic Compute Cloud. The free Anaconda Python distribution is used to handle all the packages needed and some versions of MIST are available for download under General Public License.

Reproducibility:

The plots in the poster were created using the script ExploreOptimizationResults_2019_02_24.py on Windows 10 environment with bokeh 1.0.4 holoviews 1.11.2 on Python 2.7.14 64 bit based on simulation results executed on a 64 core compute server with Ubuntu and stored in: MIST_RefModel_2019_02_18_OPTIMIZE.zip This poster can be accessed and reproduced by code accessible through the QR Images on the left.

Acknowledgments:

Thanks to Deanna J. M. Isaman who first introduced me to the the idea of accumulating knowledge from clinical trial summary data. Many thanks to Philipp Rudiger who published open source code that assisted in visualization that proved very useful. Many thanks to James Bednar who introduced holoviews that allows easy visualization of this data and to Jean-Luc Stevens for creating holoviews.
The author has no affiliation to ClinicalTrials.Gov which is an NIH project and just used here for data extraction.

Selected Publications:

Previous MSM/IMAG Posters:

Clinical Unit Mapping for Standardization of ClinicalTrials.Gov

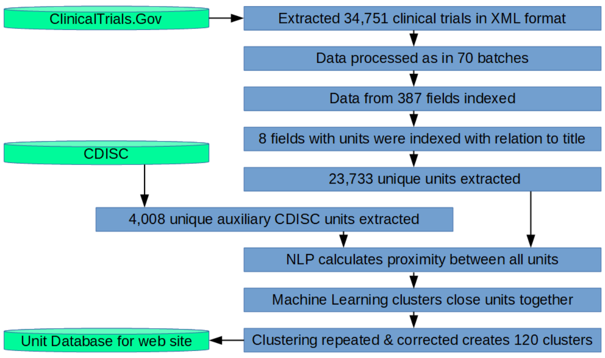
by: Jacob Barhak and Joshua Schertz

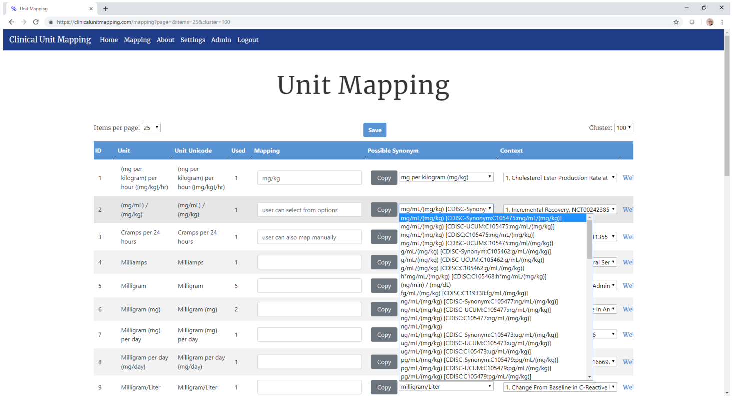
2019 IMAG Multiscale Modeling Consortium (MSM) Meeting 6-7 March 2019

Units in ClinicalTrials.Gov are not standardized, ClincialUnitMapping.com provides a solution

Abstract:

ClinicalTrials.Gov now accumulates information from over quarter of a million trials with over 10% recording trial results. It is now a U.S. Law to upload many clinical trials to this fast growing database. Data from this database can be extracted in XML format and used for modeling. The Reference Model, for example, extracts population baseline statistics and trial results. However, the database is based on textual input and although scrutinized by humans, it is currently designed for human comprehension rather than computer comprehension. Specifically, non standardized units prevent machine comprehending associated numbers. On 7 Feb 2019 all 34,751 trials with results were downloaded and unit fields were indexed and analyzed. There were 23,733 different units detected. This is a clear sign that standardization is required. We used some machine learning and Natural Language Processing algorithms to organize the data for easier processing by humans. We then created the web site: ClinicalUnitMapping.com to help standardize the units so that many models can process data in this valuable database. If units are standardized, the valuable numerical data in this database can become machine comprehensible.


ClinicalTrials.Gov
Extracted 34,751 clinical trials in XML format
Data processed as in 70 batches
Data from 387 fields indexed
8 fields with units were indexed with relation to title
CDISC
23,733 unique units extracted
4,008 unique auxiliary CDISC units extracted
NLP calculates proximity between all units
Machine Learning clusters close units together
Unit Database for web site
Clustering repeated & corrected creates 120 clusters



Unit Mapping

Items per page: 25 | Save | Cluster: 100

ID	Unit	Unit Unicode	Used	Mapping	Possible Synonym	Context
1	(mg per kilogram) per hour ([mg/kg]/hr)	(mg per kilogram) per hour ([mg/kg]/hr)	1	mg/kg	Copy mg per kilogram (mg/kg)	1, Cholesterol Ester Production Rate at Wel
2	(mg/mL) / (mg/kg)	(mg/mL) / (mg/kg)	1	user can select from options	Copy mg/mL/(mg/kg) [CDISC-Synony	1, Incremental Recovery, NCT00242385 Wel
3	Cramps per 24 hours	Cramps per 24 hours	1	user can also map manually	Copy	1355 Wel
4	Milliamps	Milliamps	1		Copy	ral Ser Wel
5	Milligram	Milligram	5		Copy	Admin Wel
6	Milligram (mg)	Milligram (mg)	2		Copy	in An Wel
7	Milligram (mg) per day	Milligram (mg) per day	1		Copy	6 Wel
8	Milligram per day (mg/day)	Milligram per day (mg/day)	1		Copy	16669 Wel
9	Milligram/Liter	Milligram/Liter	1		Copy milligram/Liter	1, Change From Baseline in C-Reactive Wel

mg/mL/(mg/kg) [CDISC-Synonym:C105475:mg/mL/(mg/kg)]
mg/mL/(mg/kg) [CDISC-UCUM:C105475:mg/mL/(mg/kg)]
mg/mL/(mg/kg) [CDISC:C105475:mg/mL/(mg/kg)]
mg/mL/(mg/kg) [CDISC-UCUM:C105475:mg/ml/(mg/kg)]
g/mL/(mg/kg) [CDISC-Synonym:C105462:g/mL/(mg/kg)]
g/mL/(mg/kg) [CDISC-UCUM:C105462:g/mL/(mg/kg)]
g/mL/(mg/kg) [CDISC:C105462:g/mL/(mg/kg)]
h*mg/mL/(mg/kg) [CDISC:C105468:h*mg/mL/(mg/kg)]
(ng/min) / (mg/dL)
fg/mL/(mg/kg) [CDISC:C119338:fg/mL/(mg/kg)]
ng/mL/(mg/kg) [CDISC-Synonym:C105477:ng/mL/(mg/kg)]
ng/mL/(mg/kg) [CDISC-UCUM:C105477:ng/mL/(mg/kg)]
ng/mL/(mg/kg) [CDISC:C105477:ng/mL/(mg/kg)]
ng/mL/(mg/kg)
ug/mL/(mg/kg) [CDISC-Synonym:C105473:ug/mL/(mg/kg)]
ug/mL/(mg/kg) [CDISC-UCUM:C105473:ug/mL/(mg/kg)]
ug/mL/(mg/kg) [CDISC:C105473:ug/mL/(mg/kg)]
pg/mL/(mg/kg) [CDISC-Synonym:C105479:pg/mL/(mg/kg)]
pg/mL/(mg/kg) [CDISC-UCUM:C105479:pg/mL/(mg/kg)]
pg/mL/(mg/kg) [CDISC:C105479:pg/mL/(mg/kg)]

Solution Key Points:

- A web site was created to allow multiple users to view and map the units
- The web site is accessible thought ClinicalUnitMapping.com

User capabilities:

- The user sees similar units clustered together and can switch clusters
- The user can see how many times the unit is used
- The user can see the contexts associated with a unit
- The user can view clinical trials that use each unit
- The user can map a unit to other suggested units
  - Clinical Data Interchange Standards Consortium (CDISC) units are suggested for mapping
  - User can see CDISC-UCUM synonyms
  - Close units are bunched together in display
- The user can ignore suggestions and provide their own mapping

Solution Construction:

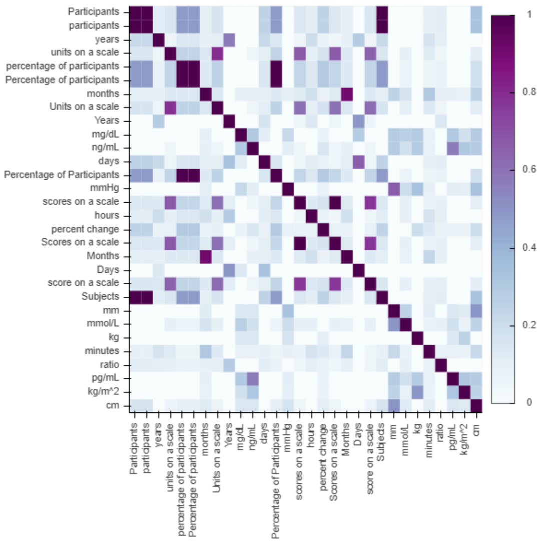
- The solution is based on python technologies that include:
  - Indexing
  - Natural Language Processing (NLP)
  - Unsupervised machine learning
  - Visualization
  - Database and web deployment

Natural Language Processing (NLP)

Units were evaluated for text proximity using two techniques developed in the Python programming language, including:

- TfidfVectorizer and cosine_similarity in scikit-learn library using 3-6 character n-grams
- difflib.SequenceMatcher method to calculate similarity ratio

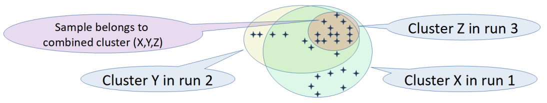
CDISC units were processed and 4008 unique units were chosen. A similarity matrix of size 23,733 x (23,733 + 4008) was constructed. The similarity to the right shows the most used units.



Machine Learning - Clustering

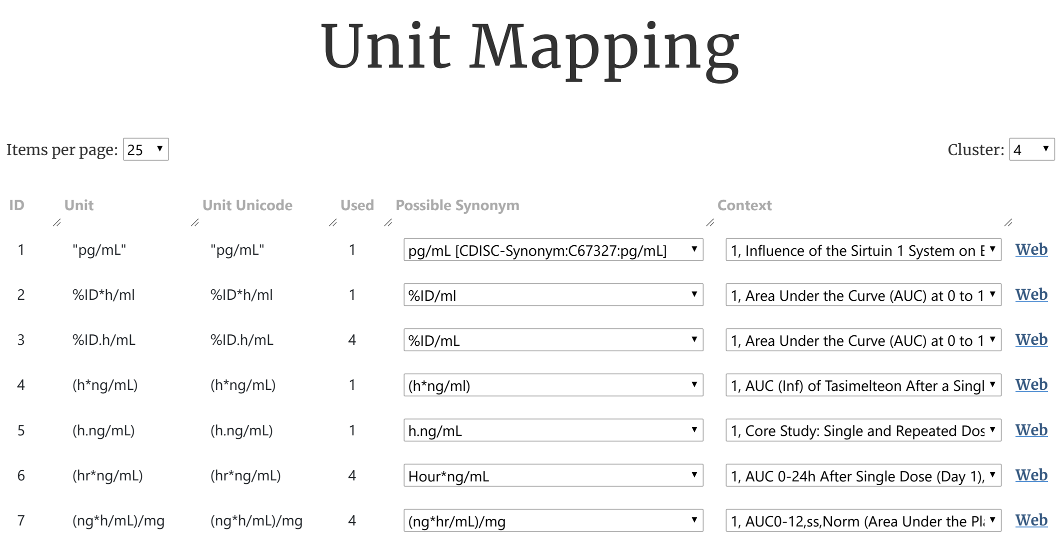
To improve user experience and allow the user to see similar units bunched together, unsupervised machine learning was applied using MiniBatchKMeans from the scikit-learn Python library.

Clustering was performed multiple times with different variations of the similarity matrix. Each unit was classified according to combined clusters considering each run - thus creating a splintering effect that ensured close units stay close. Then small clusters were eliminated by reattaching their units to the closest unit in a larger cluster. Ultimately, 129 clusters were created.



Web Site Development

The units were stored in a SQLite3 relational database. For demonstration purposes, a reduced database of only a few clusters was used as a base for the web site. The web site was developed using the Python Flask library and was deployed in a DigitalOcean instance. An administration system allows the management of multiple users, enabling a collaborative mapping effort.


Unit Mapping

Items per page: 25 | Cluster: 4

ID	Unit	Unit Unicode	Used	Possible Synonym	Context	
1	"pg/mL"	"pg/mL"	1	pg/mL [CDISC-Synonym:C67327:pg/mL]	1, Influence of the Sirtuin 1 System on E	Web
2	%ID*h/ml	%ID*h/ml	1	%ID/ml	1, Area Under the Curve (AUC) at 0 to 1	Web
3	%ID.h/mL	%ID.h/mL	4	%ID/mL	1, Area Under the Curve (AUC) at 0 to 1	Web
4	(h*ng/mL)	(h*ng/mL)	1	(h*ng/ml)	1, AUC (Inf) of Tasimelteon After a Singl	Web
5	(h.ng/mL)	(h.ng/mL)	1	h.ng/mL	1, Core Study: Single and Repeated Dos	Web
6	(hr*ng/mL)	(hr*ng/mL)	4	Hour*ng/mL	1, AUC 0-24h After Single Dose (Day 1),	Web
7	(ng*h/mL)/mg	(ng*h/mL)/mg	4	(ng*hr/mL)/mg	1, AUC0-12,ss,Norm (Area Under the Pl	Web

Discussion and Future Efforts

The goal is to solve the unit standardization problem, so that numbers can be imported easily into computer models and automated conversion to units of choice would be easy. Currently such calculations involve manual intervention and are a source of possible error. Once data is standardized, data that is currently machine readable will become machine comprehensible. Since ClinicalTrials.Gov is the largest database of clinical trials known, tapping into the knowledge stored there has great potential, and standardizing units may be only the first step. Current efforts are to expand this unit standardization project with the intention to contribute to Unified Medical Language System (UMLS) through contribution to Clinical Data Interchange Standards Consortium (CDISC) with collaboration with Simulation Interoperability Standards Organization (SISO).

Reproducibility:

Code and data for this work are archived in the file: AnalyzeCT_2019_03_02.zip. Web site database was created using the database PartUnitsDB_2018_12_26.db that was created in a previous version of the code and data archived in the files: AnalyzeCT_GOV_Code_2019_01_16.zip , StudiesWithResults_Downloaded_2018_04_20.zip. This poster can be accessed and reproduced by code accessible through the QR Images on the left.

Acknowledgments:

- Thanks to CDISC consortium help
- Thanks to NIH persons who helped and specifically to:
  - Nick Ide from NLM ClinicalTrials.Gov team on advice to process the site
  - Erin E Muhlbradt from NCI for advice on CDISC unit data

Selected Publications: